Genomic Science Program

Overview of Research Projects and Activities Underpinning Development of the DOE Systems Biology Knowledgebase

The Office of Biological and Environmental Research (BER) within the U.S. Department of Energy's (DOE) Office of Science advances world-class biological and environmental research and provides scientific facilities to support DOE missions in scientific discovery and innovation, energy security, and environmental responsibility. As a leader in systems biology, BER's Genomic Science program supports scientific research that seeks to achieve a predictive understanding of microbial and plant systems relevant to DOE missions (genomicscience.energy.gov). By revealing the genetic blueprints and fundamental principles that control the biological functions of these systems, the Genomic Science program advances the foundational knowledge underlying biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments.

Driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies, systems biology is demanding a computational environment for comparing and integrating large, heterogeneous datasets and using this information to achieve the ultimate goal of developing predictive models of biological systems. To serve the research community and meet these data-intensive computing needs, the Genomic Science program is developing the DOE Systems Biology Knowledgebase (Kbase; see Fig. 1, below). A knowledgebase is a cyberinfrastructure consisting of a collection of data, organizational methods, standards, analysis tools, and interfaces representing a dynamic body of knowledge.

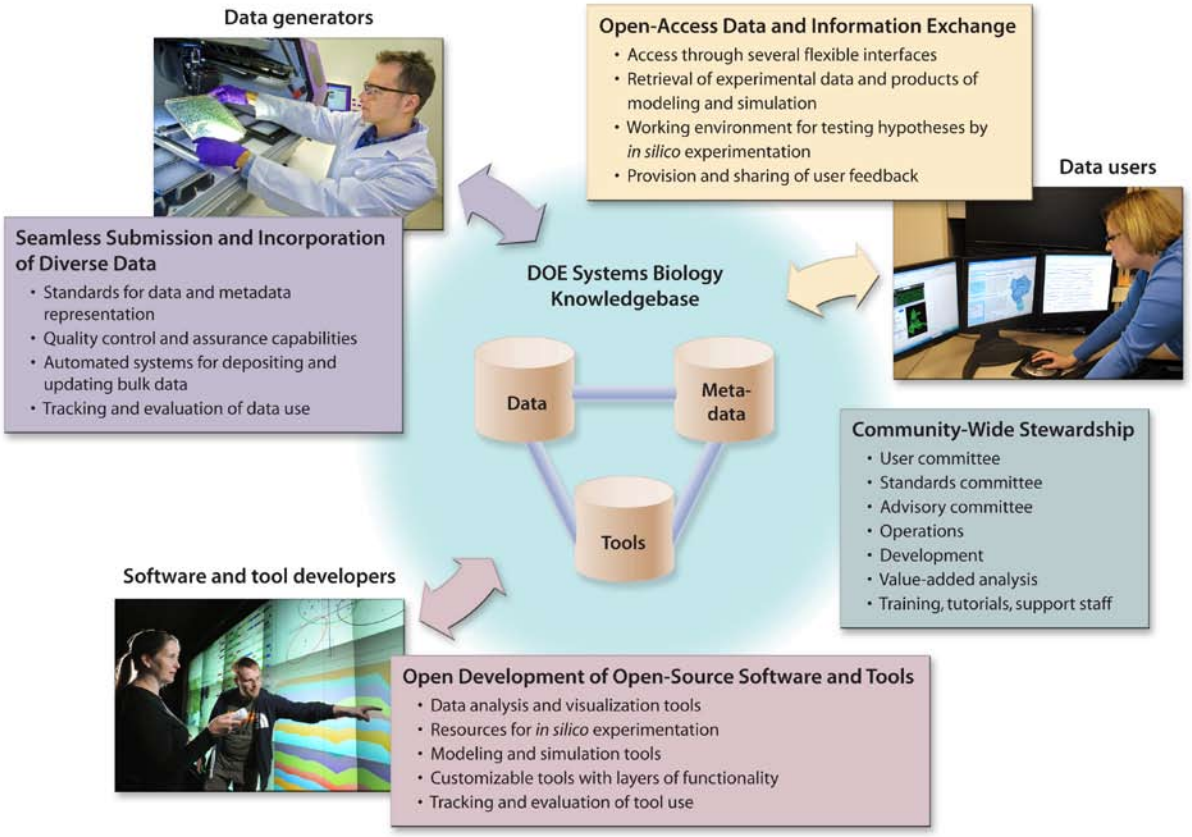


Fig. 1. Attributes and Communities Envisioned for Kbase.

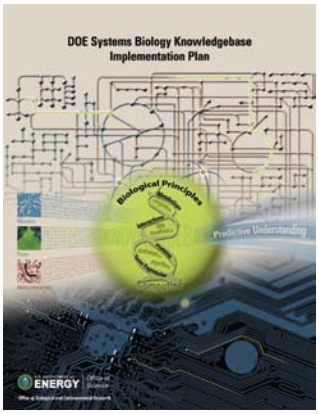
As an open, computational environment for sharing and integrating diverse biological data types, accessing and developing software for data analysis, and providing resources for modeling and simulation, Kbase supports open community science. It will differ from current informatics efforts by bringing together research products from many different projects and laboratories to create a comprehensive cyberinfrastructure focused on DOE scientific objectives in microbial, plant, and metacommunity (complex communities of organisms) research. By democratizing access to data and computational resources, Kbase will enable any laboratory or project, regardless of size, to participate in a transformative community-wide effort for advancing systems biology and accelerating the pace toward predictive biology.

This document describes Kbase development efforts carried out during the past year and summarizes current research. Specifically discussed are:

- **DOE Systems Biology Knowledgebase R&D project.** Completed in September 2010, this effort included five pilot projects and resulted in the DOE Systems Biology Knowledgebase Implementation Plan (see this page).
- **Research projects awarded in 2010 to develop computational biology and bioinformatic methods enabling Kbase.** Descriptions of the 11 funded projects (led by universities or research institutions) begin on p. 4.

Project Funded by Recovery Act Culminates in Knowledgebase Implementation Plan

In 2009, funding provided by the American Recovery and Reinvestment Act (ARRA) was used to launch the year-long DOE Systems Biology Knowledgebase R&D project. This project consisted of research and development efforts to support the conceptual design and implementation planning necessary to develop Kbase. These efforts included a series of planning workshops that brought together the systems biology and computer science communities as well as five pilot projects aimed at identifying computational problems and solutions in the context of Kbase. Together, these workshops and pilot projects informed the scientific objectives, software requirements, and design approaches detailed in the *DOE Systems Biology Knowledgebase Implementation Plan*, the final product of the R&D project. The implementation plan is a roadmap for creating Kbase and is available at two websites:



- genomicscience.energy.gov/compbio/kbase_plan/
- www.systemsbiologyknowledgebase.org

It articulates the scope and plans necessary to begin the Kbase effort and outlines a strategy for Kbase support of key research objectives in the microbial, plant, and metacommunity sciences. The report also describes the tasks, timelines, and plan for establishing Kbase's underlying infrastructure and discusses other project components such as architecture, governance, and project management.

Summary of Kbase Pilot Projects Funded by ARRA

In addition to informing and contributing to the implementation plan, the pilot projects developed software prototypes in conjunction with ARRA efforts in cloud computing funded by the DOE Office of Advanced Scientific Computing Research. These prototypes can be further integrated into the initial development of Kbase. The pilots also identified risks that must be mitigated as Kbase develops and demonstrated that such projects will be valuable in the future, especially to investigate high-risk alternatives. Descriptions of each pilot project follow.

Exploring Architecture Options for Workflows in a Federated, Cloud-Based Systems Biology Knowledgebase

- **Principal Investigator:** Ian Gorton (Pacific Northwest National Laboratory)

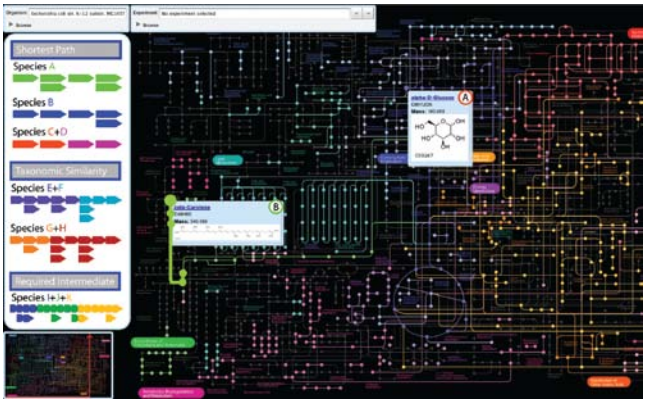
This project involved investigating available mechanisms for storing and accessing biological data in a cloud computing environment and evaluating access to large archives of omic data using a cloud architecture to provide "Data As A Service." A use case scenario to identify and curate published genome annotations was established, and investigators implemented this workflow using a federated, cloud architecture, as proposed for Kbase.

Developing Design Requirements and Prototypes of Workflows in the DOE Systems Biology Knowledgebase to Support Engineering of Metabolic Pathways

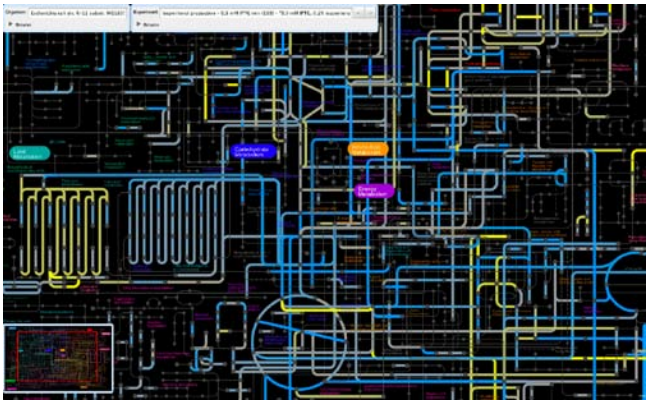
- **Principal Investigator:** Adam Arkin (Lawrence Berkeley National Laboratory)

This project designed and implemented workflows for metabolic reconstruction within MicrobesOnline, a web portal for comparative and functional genomic analyses. Investigators began developing interfaces for navigating metabolic networks and experimental functional "omics" data using the Google-like Application for Metabolic Maps or GLAMM (see Fig. 2, below). GLAMM suggests pathways that may offer routes for retrosynthesis (e.g., how to build a pathway to convert feedstock X into chemical Y in organism Z).

Fig. 2. Screenshots of GLAMM Retrosynthesis Interface (top) and GLAMM Functional Data Overlay (bottom).



Routes from a starting metabolite to a "destination" metabolite may include retrosynthesis pathways with genes from other organisms. Genes, reactions, and metabolites are linked to MicrobesOnline.



Expression data from a single experiment is rendered using the metabolic reconstruction for *Escherichia coli* K12 MG1655. Genes are mapped to the pathways they are predicted to catalyze. Increase in expression relative to control is shown in yellow; decrease is shown in blue.

Examining Technologies for Database Management Systems that Support Computational Biology and Bioinformatics Applications

- **Principal Investigator:** Victor Markowitz (Lawrence Berkeley National Laboratory and the DOE Joint Genome Institute)

This project focused on evaluating new database management system technologies that allow efficient analysis of very large datasets. Prototypes of a large database based on the DOE JGI's Integrated Microbial Genomes (IMG) data management system were implemented using several of these technologies. Performance tests of IMG "all versus all" data were conducted in Hbase on the DOE National Energy Research Scientific Computing Center's Magellan Hadoop cluster and on a smaller departmental Hadoop cluster. Results show that distributed tabular storage has significant long-term potential for Kbase but that it is not yet ready for large-scale production use. Investigators note that Hadoop and Hbase currently are undergoing rapid development, and they anticipate that stability issues will be addressed within the next 2 years.

Porting the Existing MG-RAST Multi-User Web Application to the Cloud

- **Principal Investigator:** Folker Meyer (Argonne National Laboratory)

This project investigated the requirements for distributing data across multiple platforms to optimize computational throughput. Researchers focused on the similarity analysis stage of the MG-RAST metagenome annotation server. This stage is implemented using the National Center for Biotechnology Information's BLAST resource, and investigators determined it was a good candidate for distributed computing. The project also developed guidelines for determining how best to use cloud and *ad hoc* computational resources.

Exploring Semantic-Driven Knowledge Discovery and Integration in the Systems Biology Knowledgebase Project

- **Principal Investigator:** Kerstin Kleese van Dam (Pacific Northwest National Laboratory)

This project gathered requirements to design test scenarios for semantic services such as data annotation, publication, search, access, and integration in Kbase. Investigators developed a prototype test environment that included a collaborative, project-centric user environment and a prototype data services infrastructure to support the Kbase user environment. The project demonstrated that semantic technologies are sufficiently mature to be used in a production environment to support research.

BER Researchers Begin Developing Computational Methods to Enable Kbase

In addition to the DOE Systems Biology Knowledgebase R&D project, the Genomic Science program in 2009 issued a call for proposals for research leading to the development of new computational biology and bioinformatic methods and analytics for creating Kbase. Eleven university-led projects were awarded in response to Funding Opportunity Announcement DE-FOA-0000143: *Computational Biology and Bioinformatic Methods to Enable a Systems Biology Knowledgebase*. Together, these projects seek to develop new methods to (1) integrate multiple data types, (2) infer and curate genomic and metagenomic functional annotations, (3) couple multiple cellular pathways and processes, and (4) model whole cellular processes.

Summary of University-Led Projects

Enabling a Systems Biology Knowledgebase with Gaggle and Firegoose

- **Principal Investigator:** Nitin Baliga (Institute for Systems Biology)

This project will extend the existing Gaggle and Firegoose systems to develop an open-source technology that runs over the web and links desktop applications with many databases and software applications. Researchers will incorporate workflows for data analysis that can be executed from this interface to other online applications. Four specific aims are to (1) provide one-click mapping of genes, proteins, and complexes across databases and species; (2) enable multiple simultaneous workflows; (3) expand sophisticated data analysis for online resources; and enhance open-source development of the Gaggle-Firegoose infrastructure. Gaggle is an open-source Java software system that integrates existing bioinformatics programs and data sources into a user-friendly, extensible environment to allow interactive exploration, visualization, and analysis of systems biology data. Firegoose is an extension to the Mozilla Firefox web browser that enables data transfer between websites and desktop tools including Gaggle.

Development of a Knowledgebase to Integrate, Analyze, Distribute, and Visualize Microbial Community Systems Biology Data

- **Principal Investigator:** Jill Banfield (University of California, Berkeley)

This project will develop a web-based knowledgebase that integrates metagenomic data with metaproteomic and metabolomic data from microbial communities. Although the knowledgebase will include several communities, an emphasis will be on microbes from acid mine drainage, a research area in which the principle investigator is experienced and has collected data. This new system will be usable by a larger scientific community in terms of layering gene sequence data with analyzed and predicted peptide sequence and metabolite data in a visual and queryable format. The general microbial research community likely will find this work useful. However, investigators also note that current applications to simple systems may pose interesting challenges when scaled to much larger communities. The project aims to develop three specific resources and capabilities: (1) a centralized database to integrate various omics datasets, (2) tools for mapping and representing proteomic and genomic datasets comprising orthologous genes in the presence of genomic variation, and (3) a metabolite atlas of the acid mine drainage microbial community.

Tools and Models for Integrating Multiple Cellular Networks

- **Principal Investigator:** Mark Gerstein (Yale University)

This application will develop computational tools to link metabolic pathways with regulatory pathways and physical (protein-protein) interaction data. This

work uses the principle investigator's methods for the ENCODE project (**Enc**yclopedia **o**f **D**NA **E**lements) and applies these to prokaryotes of interest to DOE. (ENCODE identifies all functional elements in the human genome sequence.) This project will go beyond ENCODE and ModENCODE (Model Organism ENCODE) by also developing topological analysis tools and dynamical modeling of integrated networks. The three specific aims are to (1) develop computational tools for analyzing integrated networks; (2) conduct correlative and topological analysis using these tools, in combination with other genomic information; and (3) carry out dynamical and evolutionary modeling of the integrated network.

Curation and Computational Design of Bioenergy-Related Metabolic Pathways

- **Principal Investigator:** Peter Karp (SRI International)

This project will develop an enhancement in the MetaCyc Pathway Tools aimed specifically at bioenergy-related processes. Pathway Tools are a set of metabolic pathway and enzyme tools generally created on an organism-by-organism basis. This application first will push out these tools to enable greater use in bioengineering for bioenergy-related processes and second will produce new graphical visualizations of metabolic pathways that can allow users to manipulate, rank, and visualize pathways. Two specific aims are to (1) enhance MetaCyc data and generate a bioenergy-related pathway and genome database and (2) develop computational tools for engineering metabolic pathways that satisfy specified design goals.

Computational Modeling of Fluctuations in Energy and Metabolic Pathways of Methanogenic Archaea

(Jointly funded with the DOE Office of Advanced Scientific Computing Research)

- **Principal Investigator:** Zaida Luthey-Schulten (University of Illinois, Urbana-Champaign)

This project will develop methodology and corresponding computational tools to simulate a population of microbes in response to environmental fluctuations. Aimed particularly at the methanogenic archaea *Methanosarcina* species, the work begins with genome-scale modeling of the microbe's metabolic and regulatory pathways. This method then will be integrated into a cellular modeling method that takes into account environmental fluctuations. Investigators will work in collaboration with William Metcalf's (University of Illinois, Urbana-Champaign) ongoing experimental studies on *Methanosarcina*. Specific aims are to (1) construct an integrated stochastic and systems model of *Methanosarcina,* (2) investigate how an *in silico* population of the microbe's cells respond to environmental fluctuations, and (3) validate the computational methodology and demonstrate its applicability to other biological systems.

A Systems Biology Knowledgebase: Context for Content

- **Principal Investigator:** Bernhard Palsson (University of California, San Diego)

This project will develop a portal and the computational tools to integrate multiple omics data to reconstruct transcriptional regulatory networks of microbes of interest to DOE (e.g., *Escherichia coli*, *Geobacter*, and *Thermotoga*). The data include protein binding (ChIP-chip), gene expression (microarrays and RNA-Seq), transcriptional start sites (sequencing), peptide (LC-FTICR-MS), and gene annotations. The application will also develop a formal mathematical framework for modeling transcriptional regulatory networks in these species. The framework captures gene-protein-reaction associations, condition-specific transcriptional basic unit structure, functional regulation of each transcriptional unit in the expression context, and structural constraints that govern transcription factor–promoter binding. Three specific aims for the project are to (1) develop computational tools to integrate omics data for genome annotation and transcription, (2) develop a genome-scale knowledgebase to provide operational constraints on cellular function, and (3) formulate *in silico* models to enable genome-scale queries.

Integrated Approach to Reconstruction of Regulatory Networks

- **Principal Investigator:** Dmitry Rodionov (Burnham Institute)

This project will extend research to identify regulons for regulatory network reconstruction and develop a method for comparing regulatory networks across microbial species. This will be accomplished by developing new clustering algorithms for cross-species comparisons, integrating known data and information from other resources and databases, and developing a platform for users to analyze experimental data. Specific aims of this application are to (1) develop an integrative platform for genome-scale regulon reconstruction,

(2) infer regulatory annotations for several groups of bacteria related to DOE missions, and (3) develop a knowledgebase for microbial transcriptional regulation data and analysis.

The final goal will be to develop a platform that integrates the experimental and computational data on transcriptional regulation in microbes. Another end goal is to allow any user to upload data (public or private), perform analyses with the data, and compare them to the analysis work conducted by the researcher who generated the data for a particular experiment.

An Open-Source Platform for Multiscale Spatially Distributed Simulations of Microbial Ecosystems

- **Principal Investigator:** Daniel Segrè (Boston University)

This project will develop an open-source platform for simulating microbial ecosystems. A simulation package will be developed based on a spatially distributed and time-dependent flux balance analysis program. One unique feature of this work will be the ability to bridge spatial and temporal scales, thus enabling simulation of microbial growth given environmental settings, including nutrient availabilities and metabolite exchange. Specific aims are to (1) modify a current dynamic flux balance analysis (dFBA) program to include spatially structured interacting metabolite dynamics of the microbial system and (2) study interactions in terms of dynamically changing colony morphology by modeling the simultaneous growth of mutualistic pairs of microbes. This work will draw on corresponding experimental data made available through project collaborators.

Phylogenomic Tools and Web Resources for the Systems Biology Knowledgebase

- **Principal Investigator:** Kimmen Sjölander (University of California, Berkeley)

This project will develop new methods to functionally annotate microbial species based on phylogenomic relationships and using the hidden Markov model (HMM) methodology based on the structural information of families of homologous genomes. The principal investigator will work collaboratively with a Harvard biologist to analyze dataset(s) containing sequence data from environmental samples of marine invertebrate-bacterial symbionts. The project also will involve collaborating with the National Institute of Advanced Industrial Science and Technology and University of Tokyo computational biologists on multispecies cooperative pathway analysis. Three primary objectives for the project are to (1) extend the PhyloFact annotation method to include new microbial data and related database information such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), PFAM, Gene Ontology (GO), experimental evidence codes, and structural information; (2) develop a new HMM algorithm to create novel gene trees; and (3) apply the PhyloFact annotation pipeline to collaborative marine microbial systems.

Development of an Extensible Computational Framework for Centralized Storage and Distributed Curation and Analysis of Genomic Data and Genome-Scale Metabolic Models

- **Principal Investigator:** Rick Stevens (University of Chicago)

This work will develop a computational framework that combines a centralized extensible database for integrating omics and sequence data with a distributed pipeline for using these data to annotate genomes and to reconstruct and analyze new genome-scale metabolic models. The proposed framework will be interfaced with the SEED. Three significant components of this interface will be enhancing the backend of SEED to support new data types and queries, integrating this into a model-building application for whole genome–scale networks (regulatory and metabolic) and developing an application programming interface (API) for Kbase to utilize this work.

Specific project objectives are (1) an improved infrastructure to enhance the framework's extensibility, accessibility, and scalability; (2) an extended database to accommodate new predicted and experimental biological data types such as microbial transcriptional regulatory networks, genome-scale metabolic models, experimental evidence (e.g., microarray data, ChIP-chip data, and equilibrium constants), eukaryote genomes, and growth phenotype data (e.g., biology array data, culture conditions, growth rates, and gene essentiality); and (3) a new API to provide remote access to the database and tools, including RAST annotation of raw genome sequences, automated reconstruction of draft genome-scale metabolic models, flux balance analysis of such models; and querying of all data.

Gene Ontology Terms and Automated Annotation for Energy-Related Microbial Genomes

- **Principal Investigators:** Brett Tyler and João Carlos Setubal (Virginia Polytechnic Institute and State University)

This effort will develop a set of GO terms for describing energy-related microbial processes. GO is one of the more widely used functional ontologies for annotating genes, and this project will address the known community gap in GO terms for microbial processes that makes GO much more relevant for human systems. Two specific aims are to (1) develop MENGO terms (ontologies for microbial energy processes) and host a series of tutorials and workshops at key meetings to inform and train microbiologists on these terms and (2) develop a database and web interface for storing and displaying these terms and microbial annotations.

Contact

Dr. Susan Gregurick
U.S. Department of Energy
Office of Biological and Environmental Research
Phone: 301.903.7672
Email: susan.gregurick@science.doe.gov

Reports

***DOE Systems Biology Knowledgebase Implementation Plan*, September 2010**

- genomicscience.energy.gov/compbio/kbase_plan/
- www.systemsbiologyknowledgebase.org

***DOE Systems Biology Knowledgebase for a New Era in Biology*, March 2009**

- genomicscience.energy.gov/compbio/

Websites

DOE Office of Science

- www.science.doe.gov

DOE Office of Biological and Environmental Research

- www.science.doe.gov/ober/

Genomic Science program

- genomicscience.energy.gov